An Overview of k-Means Clustering Algorithm

E. U. Oti[1]; J. Ikirigo[2]; P. A. Esemokumo[3]

[1,3] Department of Statistics,
Federal Polytechnic, Ekowe, Bayelsa State, Nigeria

[2]Department of Physics with Electronics,
Federal Polytechnic, Ekowe, Bayelsa State, Nigeria
E-mail: eluchcollections@gmail.com[1]

***Abstract* — This paper presents a comprehensive review of existing techniques of k-means clustering algorithms made at various times. The k-means algorithm is aimed at partitioning objects or points to be analyzed into well-separated clusters. There are different algorithms for k-means clustering of objects such as traditional k-means algorithm, standard k-means algorithm, basic k-means algorithm, and the conventional k-means algorithm, this is perhaps the most widely used version of the k-means algorithms. These algorithms use the Euclidean distance as their metric and minimum distance rule approach by assigning each data point (object) to its closest centroids.**

Keywords: *k-means, Cluster analysis, Centroid, Euclidean distance, Unsupervised classification.*

I. Introduction

Data clustering is an important topic of research and it has its applications in various fields like statistics, data mining, computer science, pattern recognition, image processing, marketing, psychiatry, etc. (Anderberg, 1973; Brohée and Helden, 2006; Everitt et al., 2011). Clustering is seen as purely a multivariate technique but it can also be applied to univariate and bivariate data. Clustering or grouping is done based on similarities or distances (Johnson and Wichern, 2002) and it is one of the best approaches of multivariate analysis and a common methodology for statistical data analysis.

Clustering (cluster analysis) was originated in anthropology by Driver and Kroeber (1932) and was introduced to psychology in 1938 (Zubin, 1938). Cattell (1949) introduced mathematical procedures for organizing objects based on observed similarity. It was not until Sokal and Sneath's (1963) publication of Principles of Numerical Taxonomy; that the clustering method gained widespread acceptances in the sciences, and motivated worldwide research on clustering methods and thereby initiated the publication of a broad range of books such as those of Fisher (1968), Tryon and Bailey (1970), Jardine and Sibson (1971), Anderberg (1973), Hartigan (1975), Späth (1980,1985), Aldenderfer and Blashfield (1984), Romesburg (1984), Fukunaga (1990), Kaufman and Rousseeuw (1990), Berkhin (2006), Mirkin (2013), etc.

K-means is the most popular clustering formulation in which the goal is to maximize the expected similarity between data items and their associated cluster centroids (Slonim et al., 2013)

The purpose of this paper is to present a comprehensive review of existing techniques of k-means clustering algorithm made at various times.

The rest of this paper is organized as follows: section 2 discussed hierarchical and partitioning clustering methods as the main group of cluster analysis. Section 3 discussed k-means clustering. Furthermore, section 4 discussed related literature on k-means clustering. Finally, section 5 is the conclusion of the paper..

.

II. Cluster Analysis

Cluster analysis or clustering is an unsupervised classification mechanism where a set of data, usually multidimensional is classified into groups (clusters) such that members of one cluster are similar to one another with respect to some predetermined criterion (Hartigan, 1975; Jain and Dubes, 1988; Mirkin, 2013).

Cluster analysis can be divided into two main groups which are based on the structure of their output namely: hierarchical non-hierarchical (Partitioning) clustering methods. Hierarchical clustering also known as hierarchical cluster analysis is an algorithm that groups similar objects into groups called clusters. The clusters are merged (agglomerative methods) or split (divisive methods) step-by-step based on the similarity measure. The results of a hierarchical clustering method entails that agglomerative and divisive methods can be displayed graphically using a tree diagram known as dendrogram.

The dendrogram shows all the steps in the hierarchical procedure which includes the similarities or distances at which clusters are merged. While partitioning clustering methods partition the data object set into clusters where every pair of object clusters is either distinct or has some members in common. Partitioning clustering begins with a starting cluster partition which is iteratively improved until a locally optimal partition is reached (Hartigan, 1975).

2.1 *K-meaans Clustering*

K-means is an iterative procedure that partition N objects into K disjoint clusters. K-means is perhaps the most widely used clustering method, and especially the best-known of the partitioning-based clustering methods that uses centroids for cluster presentation (Estivill-Castro, 2002). The quality of k-means clustering is measured through the within-cluster squared error criterion (MacQueen, 1967; Yuan and Yang, 2019; Hastie et al., 2001).

K-means algorithm is used to minimize the problem of k-means, and it has many variants which will be discussed next but to be able to use any of the k-means algorithm, the number of clusters present in the data need to be known; multiple runs or trials will be necessary to find the best number of clusters. There is no best k-means algorithm, as the tendency of generating global optimum depends on the characteristics of the data set, the size and also the number of variables in the cases. The k-means clustering methods have two phases of iteration namely: the assignment or initialization phase which involves an iterative process where each data point is assigned to its nearest centroid using Euclidean metric; the next is the centroid update phase, where clusters centroids are updated given the partition obtained by the previous phase. The iterative process stops when no data point change clusters or some maximum number of iterations is reached (Slonim et al., 2013).

Forgy (1965) proposed a batch algorithm called the traditional k-means algorithm; the algorithm is based on the minimization of the average squared Euclidean distance between the data points and the cluster's center known as centroid, where centroid is the center of a geometric object and it is seen as a generalization of the mean. The Forgy's algorithm start by choosing the number of cluster k representing the cluster centers, it then assigns data point of the data set to the cluster having the closest centroid, update new centroids for each cluster by averaging the data points or objects belonging to the cluster, if there is no change in the cluster center, then the iteration stops.

Lloyd (1982) proposed the standard k-means algorithm which is also a batch algorithm, the difference between Forgy's algorithm and Lloyd's algorithm is that Forgy's algorithm treats data distribution as continuous while Lloyd's algorithm treats data distribution as discrete case.

MacQueen (1967) proposed the basic k-means algorithm which is an online algorithm, the algorithm is similar to Forgy's and Lloyd's algorithm when it comes to the initialization process but differs from the two algorithms when it comes to the update process. During the update of MacQueen's algorithm, the centroids are updated by re-calculating the points any time there is a change in the centroid, and when each points is currently assigned to the cluster with the nearest centroid, the process stops.

Hartigan and Wong (1979) proposed a conventional k-means algorithm which is a non-Forgy (or non-Lloyd) heuristic that updates cluster centers considering each points, rather than after each pass over the entire data set. This algorithm searches for the partition of data space with locally optimal within-cluster sum of squares of errors (SSE); which means that it may assign a point to another subspace, even if it currently belongs to the subspace of the closest centroid. If the centroid has been updated for each data point included, the within-cluster sum of squares for each data point if included in another cluster is calculated. If one of the cluster sum of squares (SSE 2 in the equation below, for all $i \neq 1$), the point is assigned to this new cluster

$$SSE\ 2 = \frac{N_i \sum_j^k \|x_{ij} - c_i\|^2}{N_i - 1} < SSE\ 1 = \frac{N_1 \sum_j^k \|x_{ij} - c_i\|^2}{N_1 - 1}$$

Where N_i is the number of points included in cluster k, x_{ij} is the jth point in the $ith\ cluster$ and c_i is the ith point in the cluster center. The iteration continuous until no point changes cluster.

.

IV. Related Literature

Jancey (1966) proposed a variant which is a modification for the Forgy's k-means algorithm (cf. Anderberg, 1973) which is expected to accelerate convergence and inferior local minima. In this variant, the new cluster center is not the mean of the old and added points, but the new center is updated by reflecting the old center through the mean of the new cluster.

In order to avoid poor local solutions, a number of genetic algorithm based methods have been developed (Krishna and Murty, 1999; Bandyopadhyay and Maulik, 2002). Likas et al. (2003) developed the global k-means clustering algorithm which is a deterministic and incremental global optimization method. It is also independent on any initial parameters and employs k-means procedure as a local search procedure, since the

exhaustive global k-means method is computationally expensive.

Faber (1994) proposed a variant of the Lloyd's k-means algorithm called the continuous k-means algorithm. The reference points in the continuous k-means algorithm are chosen as a random sample from the whole population of the data point while in the standard k-means algorithm, the initial reference points are chosen more or less arbitrarily. During the update process, the continuous k-means algorithm examines only a random sample of the data points while the standard k-means algorithm examines all of the data set in sequence. If the data set is very large and the sample is a representative of the data set, then the continuous k-means algorithm should converge much faster than the algorithm that examines every point in sequence.

Kanungo et al. (2002) presented a simple and efficient implementation of Lloyd's k-means clustering algorithm which they called the filtering algorithm. The filtering algorithm is easy to implement which requires a kd-tree (cf. Bentley, 1975) as the only major data structure. A kd-tree is a binary tree, which represents a hierarchical sub-division of the point set's bounding box using axis aligned splitting hyperplanes. Each node of the kd-tree is associated with a closed box, called a cell. The root's cell is the bounding box of the point set. If the cell contains at most one point (or, more generally, fewer than some small constant), then it is declared to be a leaf. Otherwise, the root's cell is splitting into two hyper-rectangles by an axis orthogonal hyperplane. The points of the cell are then partitioned to one side or the other of this hyperplane. The resulting sub-cells are the children of the original cell, thus leading to a binary tree structure.

Bagirov and Mardaneh (2006) proposed a new variant of the global k-means algorithm which is known as the modified global k-means (MGKM) algorithm because it is said to be effective for solving clustering problems in gene expression data sets. In their algorithm, a starting point for the kth cluster center is computed by minimizing the so-called auxiliary cluster function. The effectiveness of this algorithm highly depends on its starting point. The algorithm computes clusters incrementally and to compute k-partition of a data set, it uses $k-1$ cluster centers.

Nazeer and Sebastpan (2009) discussed in their paper about one major drawback of k-means algorithm, they proposed an enhanced method that deals with improving the accuracy and efficiency of k-means algorithm. Both the phases of the original k-means algorithm were modified. The initial centroids are determined systematically so as to produce clusters with better accuracy in the first phase. The second phase makes use of a variant of the clustering method discussed in Fahim et al. (2006). It starts by forming the initial clusters based on the relative distance of each data point from the initial centroids. These clusters are subsequently fine-tuned by using a heuristic approach there by improving the efficiency.

Huang et al. (2005) proposed a k-means type clustering algorithm that can automatically calculate variable weights and it is referred to as weighted-k-means (W-K-Means). The weighted-k-means adds a new step to the basic k-means algorithm to iteratively update the variable weights based on the current partition of the data and also a formula for weights calculation was proposed as well. The variable weights produced by the proposed weighted-k-means algorithm measured the importance of variables in clustering and can be used in variable selection in data mining applications where large and complex real data are often involved.

Amorim (2012) proposed the constrained Minkowski Weighted K-Means algorithm which calculates cluster specific feature weights that can be interpreted as features rescaling factors. Naturally, the Minkowski weighted k-means (MWK-Means) algorithm requires a Minkowski exponent, p, which can be approximated via semi-supervised learning (Amorim and Mirkin, 2012). Weight w_{kw} was introduced, which depends on both cluster k, and feature v, allowing a given feature v, to have different weights at different cluster k; also, the use of the Minkowski distance to the power of p was introduced, analogous to the Euclidean squared distance $d_p(y_i, c_k) = \sum_{v=1}^{v} W_{kw}^{p} |y_{iv} - c_{kw}|^p$

where v represents the features and p is the Minkowski exponent, w_{kv}^{p} is the weight variable to take into account the Minkowski exponent p. Wagstaff et al. (2001) introduced constrained clustering k-means which makes use of limited amount of background knowledge by applying pairwise must-link and cannot-link rules to entities and likewise is the Minkowski weighted k-means.

.

V. Conclusion

In this paper, we have reviewed existing techniques in k-means clustering. This work shows that there are several variants of k-means clustering algorithms from sixties to recent times which have addressed some drawback of k-means algorithms.

In the future, we will look at the computational time complexity of some of the variants of k-means clustering algorithms and its analysis in terms of relative accuracy and efficiency.

Acknowledgements
The authors wish to thank the referees for their worthwhile comments and suggestions. This research was sponsored by Nigerian Tertiary Education Trust Fund (TETFUND).

References

Aldenderfer, M. and Blashfield, R. (1984). Cluster Analysis. Newbury Park, CA: Sage Publications.

Amorim, R. C. (2012). Constrained clustering with Minkowski weighted k-means. Proceedings of the 13th IEEE International Symposium on Computational Intelligence and Informatics, pp.13-17.

Amorim, R. C. and Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. Pattern Recognition, Vol. 45:1061-1075.

Anderberg, M. R. (1973). Cluster Analysis for Applications. New York: Academic Press.

Bagirov, A. M. and Mardaneh, K. (2006). Modified global k-means algorithm for clustering in gene expression datasets. Conference Proceedings Workshop on Intelligent Systems for Bioinformatics, Vol. 73: 23-28.

Bandyopadhyay, S. and Maulik, U. (2002). An evolutionary technique based on k-means algorithm for optimal clustering in R^N, Information Science, 146, pp. 221-237.

Bentley, J. L. (1975). Multidimensional binary Search Trees used for associative searching. Communication of the ACM, 18(9), 509-517.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques, Inc: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data, Springer, Berlin, Heidelberg.

Brohée, S. and Helden, J. V. (2006). Evaluation of Clustering Algorithms for Protein-Protein Interaction Networks, BMC Bioinformatics, Vol.7, pp. 1-19.

Cattell, R. (1949). r_p and other coefficients of pattern similarity. Psychometrika, 14(4):279-298.

Driver, H. E. and Kroeber, A. L. (1932). "Quantitative Expression of Cultural Relationships," University of California Publications of American Archaeology and Ethnology, 31(4), pp. 211-256.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. ACM SIGKDD Explorations Newsletter, 4(1):65-75.

Everitt, B., Landau, S., Leese, M., and Stajl, D. (2011). Cluster analysis, 5thedition, John Wiley and Sons.

Faber, V. (1994). Clustering and the continuous k-means algorithm. Los Alamos Science, 22:138-144.

Fahim, A. M., Salem, A. M., Torkey, F. A., and Ramadan, M. A. (2006). "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University. 10(7): 1626-1633.

Fisher, W. D. (1968). Clustering and Aggregation in Economics: Johns Hopkins Press, Baltimore, Maryland.

Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, 2nd edition, Academic Press.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classification. Biometrics, Vol. 21, pp. 768-769.

Hartigan, J. A. (1975). Clustering Algorithms. John Wiley & Sons. Inc., New York.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136.A k-means clustering algorithm. Applied Statistics. 28, 100-108.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). The elements of Statistical learning: Data mining, inference and prediction, Springer-Verlag.

Huang, J. Z., Ng, M. K., Rong, H. and Li, Z. (2005). "Automated variable weighting in k-means type clustering". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 27(5), pp. 657-668.

Jain, A. K. and Dubes, R. (1988). Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall.

Jancey, R. C. (1966). Multidimensional Group Analysis, Australian Journal of Botany, 14(1), pp. 127-130.

Jardine, N. and Sibson, R. (1971). Mathematical Taxonomy. John Wiley and Sons, Ltd, Chichester.

Johnson, R. A. and Wichern, D. W. (2002). Applied Multivariate Statistical Analysis.5th Edition, Eaglewood Cliffs, NJ: Prentice- Hall.

Kanungo, T., Mount, D., Netanyahu, N.,Piatko, C.,Silverman, R., and Wu, A. (2002). An efficient k-means clustering algorithm: Analysis and Implementation .IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881-892.

Kaufman, L. and Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc.

Krishna, K. and Murty, M. (1999). Genetic K-Means algorithm. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 29(3): 433-439.

Likas, A., Vlassis, N. and Verbeek, J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451-461.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transaction on information Theory, 28(2), 129-137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp.281-297. Berkeley, CA: University of California Press.

Mirkin, B. (2013). Clustering: A Data Recovery Approach, Second Edition (Chapman and Hall/CRCComputer Science and Data Analysis).

Nazeer, K. A. A. and Sebastian, M. P. (2009). Improving the accuracy and efficiency of the k-means clustering algorithm, Proceedings of the World Congress on Engineering, Vol. 1, pp. 1-3.

Romesburg, C. (1984). Cluster Analysis for Researchers. London: Wadesworth.

Slonim, N., Aharoni, E. and Crammer, K. (2013). Hartigan's K-Means Versus Lloyd's K-Means-Is It Time for a Change? Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 1677-1684.

Sokal, R. and Sneath, P. (1963). Principles of Numerical Taxonomy. San Francisco: California.

Spáth, H. (1980). Cluster Analysis Algorithms. West Sussex, UK: Ellis Horwood Limited.

Spáth, H. (1985). Cluster Dissection and Analysis, Ellis Horwood, Chichester.

Tryon, R. C. and Bailey, D. C. (1970). Cluster Analysis. McGraw Hill, New York.

Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. (2001). Constrained k-means clustering with background knowledge. In Proceedings of the 8th International Conference on Machine Learning. pp. 577-584.

Yuan, C. and Yang, H. (2019). Research on k-value selection method of k-means clustering algorithm. Multidisciplinary Scientific Journal, 2(2), 226-235.

Zubin, J. (1938). "A technique for measuring like-mindedness".The Journal of Abnormal and Social Psychology. 33(4), 508-516.